DETAILED REPORT

FIRST REVIEW

COVID-19 DATA ANALYSIS AND FORECASTING

GUIDE SIGNATURE

SUBMITTED BY

REAHAAN SHERIFF I

2019202045

MCA – R 3YRS

COVID-19 DATA ANALYSIS AND FORECASTING

A PROJECT REPORT

Submitted by

REAHAAN SHERIFF I – 2019202045

A report of the project submitted to the Faculty of

INFORMATION SCIENCE AND TECHNOLOGY

in partial fulfillment

for the award of the degree

of

MASTER OF COMPUTER APPLICATION

DEPARTMENT OF INFORMATION SCIENCE AND TECHNOLOGY,

COLLEGE OF ENGINEERING, GUINDY

ANNA UNIVERSITY CHENNAI 600 025

MAY, 2022

ANNA UNIVERSITY

CHENNAI - 600 025

BONAFIDE CERTIFICATE

Certified that this project report titled COVID-19 DATA ANLYSIS AND FORECASTING is the bonafide work of REAHAAN SHERIFF I who carried out project work under my supervision. Certified further that to the best of my knowledge and belief, the work reported herein does not form part of any other thesis or dissertation based on which a degree or an award was conferred on an earlier occasion on this or any other candidate.

PLACE: CHENNAI

DATE: 10-05-2022

PROJECT GUIDE
Ms. P.S APIRAJITHA
TEACHING FELLOW
DEPARTMENT OF IST
ANNA UNIVERSITY
CHENNAI 600025

DR.S.SRIDHAR

HEAD OF THE DEPARTMENT
DEPARTMENT OF INFORMATION SCIENCE AND TECHNOLOGY
COLLEGE OF ENGINEERING, GUINDY
ANNA UNIVERSITY
CHENNAI 600025

ABSTRACT

COVID-19 has sparked a worldwide pandemic, with the number of infected cases and deaths rising on a regular basis. Along with recent advances in soft computing technology, researchers are now actively developing and enhancing different mathematical and machine-learning algorithms to forecast the future trend of this pandemic. Thus, if we can accurately forecast the trend of cases globally, the spread of the pandemic can be controlled. In this project, a LSTM model will be used on a time-series dataset to forecast the cases of COVID-19 in future.

Machine learning (ML) is a category of an algorithm that allows software applications to become more accurate in predicting outcomes without being explicitly programmed. The basic premise of machine learning is to build algorithms that can receive input data and use statistical analysis to predict an output while updating outputs as new data becomes available.

Long Short Term Memory is a kind of recurrent neural network. In RNN output from the last step is fed as input in the current step. It tackled the problem of long-term dependencies of RNN in which the RNN cannot predict the word stored in the long-term memory but can give more accurate predictions from the recent information. As the gap length increases RNN does not give an efficient performance. LSTM can by default retain the information for a long period of time. It is used for processing, predicting, and classifying on the basis of time-series data.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Machine learning (ML) is a category of an algorithm that allows software applications to become more accurate in predicting outcomes without being explicitly programmed. The basic premise of machine learning is to build algorithms that can receive input data and use statistical analysis to predict an output while updating outputs as new data becomes available.

Long Short Term Memory is a kind of recurrent neural network. In RNN output from the last step is fed as input in the current step. It tackled the problem of long-term dependencies of RNN in which the RNN cannot predict the word stored in the long-term memory but can give more accurate predictions from the recent information. As the gap length increases RNN does not give an efficient performance. LSTM can by default retain the information for a long period of time. It is used for processing, predicting, and classifying on the basis of time-series data.

1.1 MOTIVATION AND OBJECTIVE:

To predict the cases and peak case date for COVID-19 in India as accurately as possible. We will be using LSTM networks.

1.2 IMPLEMENTATION PLATFORM / FRAMEWORK:

Google COLAB used to implement this project using LSTM machine learning algorithm.

Libraries used,

1. Pandas,
2. Numpy,
3. Seaborn,
4. Plotly,
5. Tensorflow
6. Matplotlib.

CHAPTER 2

LITERATURE REVIEW

2.1 Coronavirus disease (COVID-19) cases analysis using machine-learning applications

Today world thinks about coronavirus disease that which means all even this pandemic disease is not unique. The purpose of this study is to detect the role of machine-learning applications and algorithms in investigating and various purposes that deals with COVID-19. Review of the studies that had been published during 2020 and were related to this topic by seeking in Science Direct, Springer, Hindawi, and MDPI using COVID-19, machine learning, supervised learning, and unsupervised learning as keywords. The total articles obtained were 16,306 overall but after limitation; only 14 researches of these articles were included in this study. Our findings show that machine learning can produce an important role in COVID-19 investigations, prediction, and discrimination. In conclusion, machine learning can be involved in the health provider programs and plans to assess and triage the COVID-19 cases. Supervised learning showed better results than other Unsupervised learning algorithms by having 92.9% testing accuracy. In the future recurrent supervised learning can be utilized for superior accuracy.

CHAPTER 3

SYSTEM DESIGN

This chapter consists of the system design of the project with the overall architecture and the description of the modules used in the project

3.1 OVERALL ARCHITECTURE:

The overall system architecture of the proposed system is shown in Figure 3.1. For analysis the time series dataset will be taken and it will be pre-processed. That pre-processed data will be splitted into training and testing. The trained dataset will be applied to the LSTM model and that model will be used for evaluating the test dataset. After the results the model will be used for future forecasting to forecast the upcoming cases and the evaluation of those results will be done in RMSE, MSE and MAE metrics.

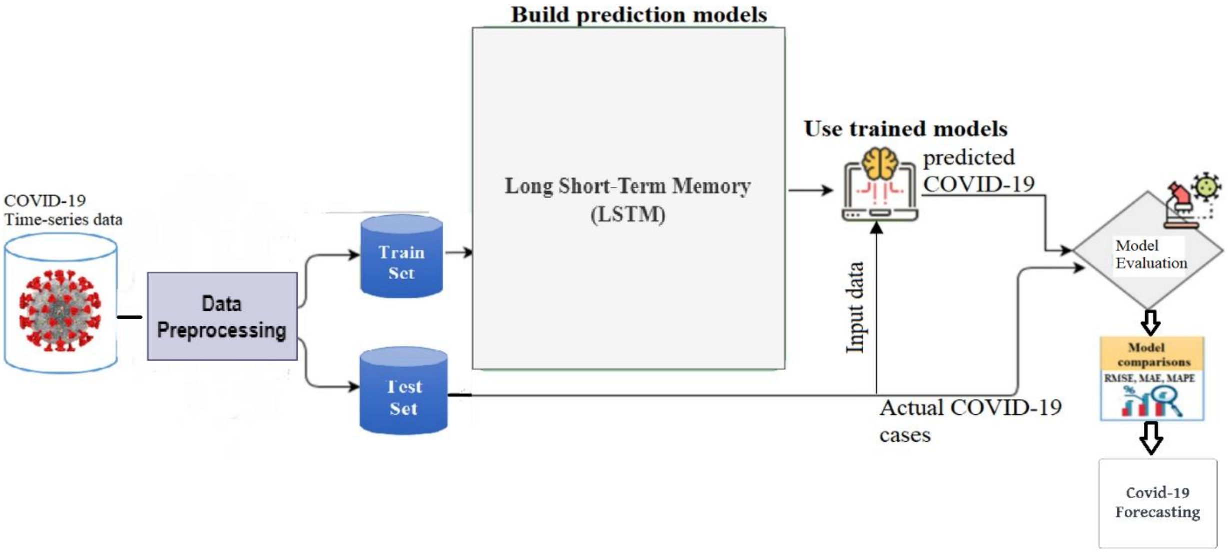


Figure 3.1: Architecture Diagram

3.2 MODULES DESCRIPTION:

3.2.1 Data pre-processing

Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model.

3.2.2 Building model

We will built the model with the help of LSTM. The model has an input layer followed by three LSTM layers. The LSTM layers contain Dropout as 0.5 to prevent overfitting in the model. The output layer consists of a dense layer with 1 neuron with activation as relu. We will predict the number of Corona cases, so our output will be a positive number $(0, \infty)$.

3.2.3 Training model

To train the model, we will take out training data (80%) and used 20% of it as validation data. To lower the learning rate of our model we will use reducelronplateau in the model. Training the model with n epochs.

3.2.4 Model evaluation

It estimates how well (or how bad) the model is, in terms of its ability in mapping the relationship between X (a feature, or independent variable, or predictor variable) and Y (the target, or dependent variable, or response variable).

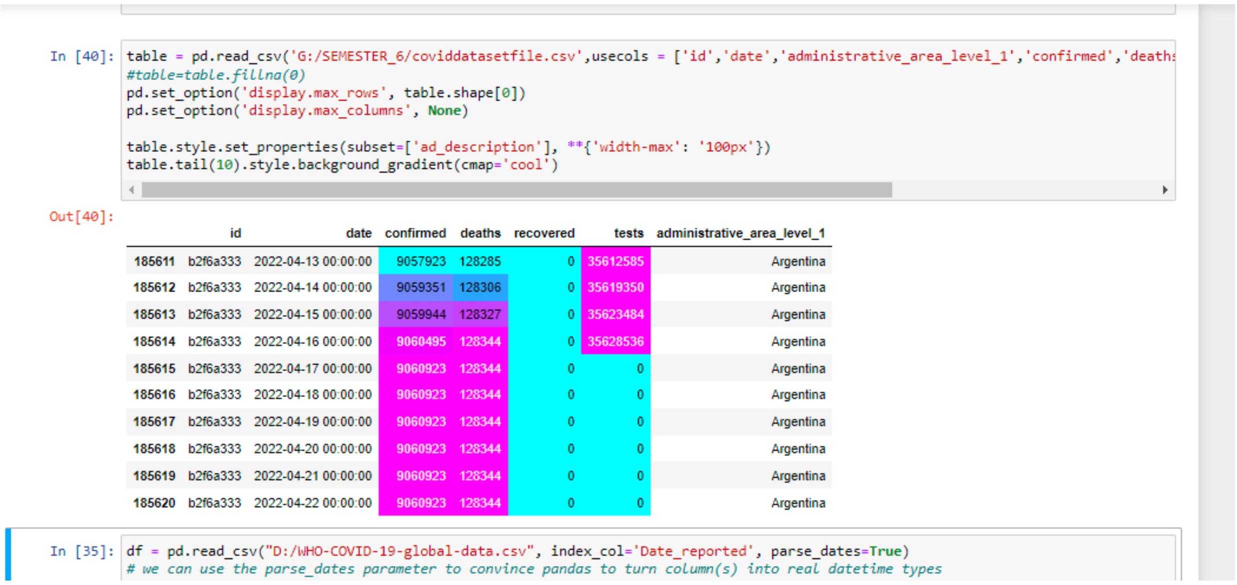
3.2.5 Visualization and forecasting

In order to see the prediction and accuracy, first, we predicted the output of our x_test data. This was the output that we got from the test data. To accurately plot the values, we needed to bring our prediction and y_test data back to the original bounds of the data. In the end, we plotted a graph between the actual COVID-19 cases compared to our predicted COVID-19 cases to see the overall accuracy of our model

CHAPTER 4

MODULE-BASED OUTPUT SCREENSHOTS

Dataset:



```
In [40]: table = pd.read_csv('G:/SEMESTER_6/coviddatasetfile.csv',usecols = ['id','date','administrative_area_level_1','confirmed','deaths
         #table=table.fillna(0)
         pd.set_option('display.max_rows', table.shape[0])
         pd.set_option('display.max_columns', None)

         table.style.set_properties(subset=['ad_description'], **{'width-max': '100px'})
         table.tail(10).style.background_gradient(cmap='cool')
```

Out[40]:

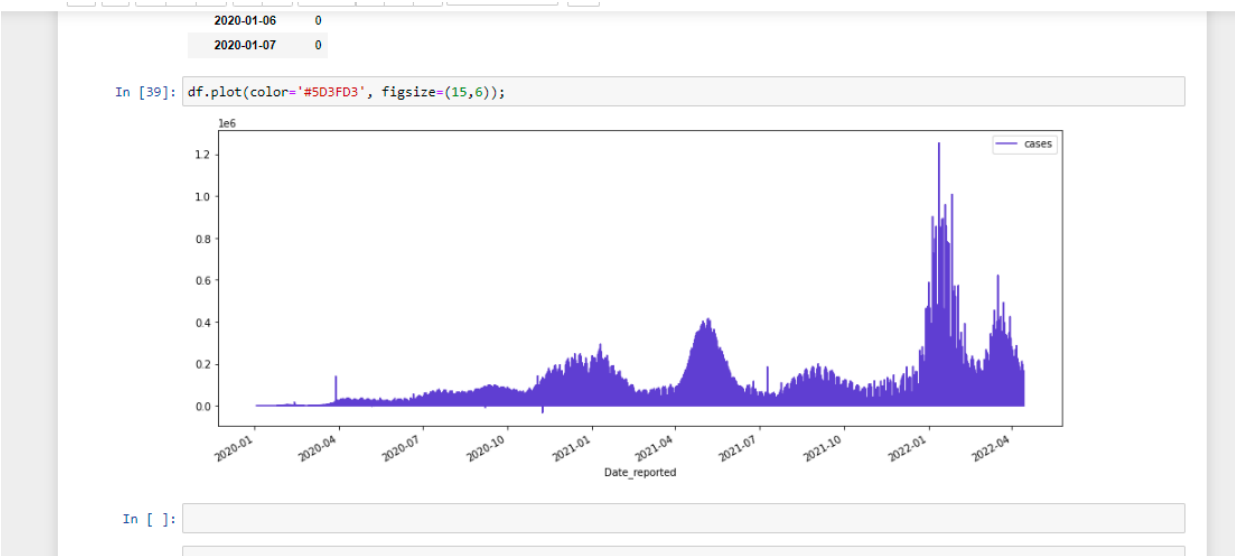
	id	date	confirmed	deaths	recovered	tests	administrative_area_level_1
185611	b2f6a333	2022-04-13 00:00:00	9057923	128285	0	35612585	Argentina
185612	b2f6a333	2022-04-14 00:00:00	9059351	128306	0	35619350	Argentina
185613	b2f6a333	2022-04-15 00:00:00	9059944	128327	0	35623484	Argentina
185614	b2f6a333	2022-04-16 00:00:00	9060495	128344	0	35628536	Argentina
185615	b2f6a333	2022-04-17 00:00:00	9060923	128344	0	0	Argentina
185616	b2f6a333	2022-04-18 00:00:00	9060923	128344	0	0	Argentina
185617	b2f6a333	2022-04-19 00:00:00	9060923	128344	0	0	Argentina
185618	b2f6a333	2022-04-20 00:00:00	9060923	128344	0	0	Argentina
185619	b2f6a333	2022-04-21 00:00:00	9060923	128344	0	0	Argentina
185620	b2f6a333	2022-04-22 00:00:00	9060923	128344	0	0	Argentina

```
In [35]: df = pd.read_csv("D:/WHO-COVID-19-global-data.csv", index_col='Date_reported', parse_dates=True)
         # we can use the parse_dates parameter to convince pandas to turn column(s) into real datetime types
```

Figure 4.1: EDA

This image displays the required data columns for training the machine learning model in Figure 4.1 home screen.

Graph:

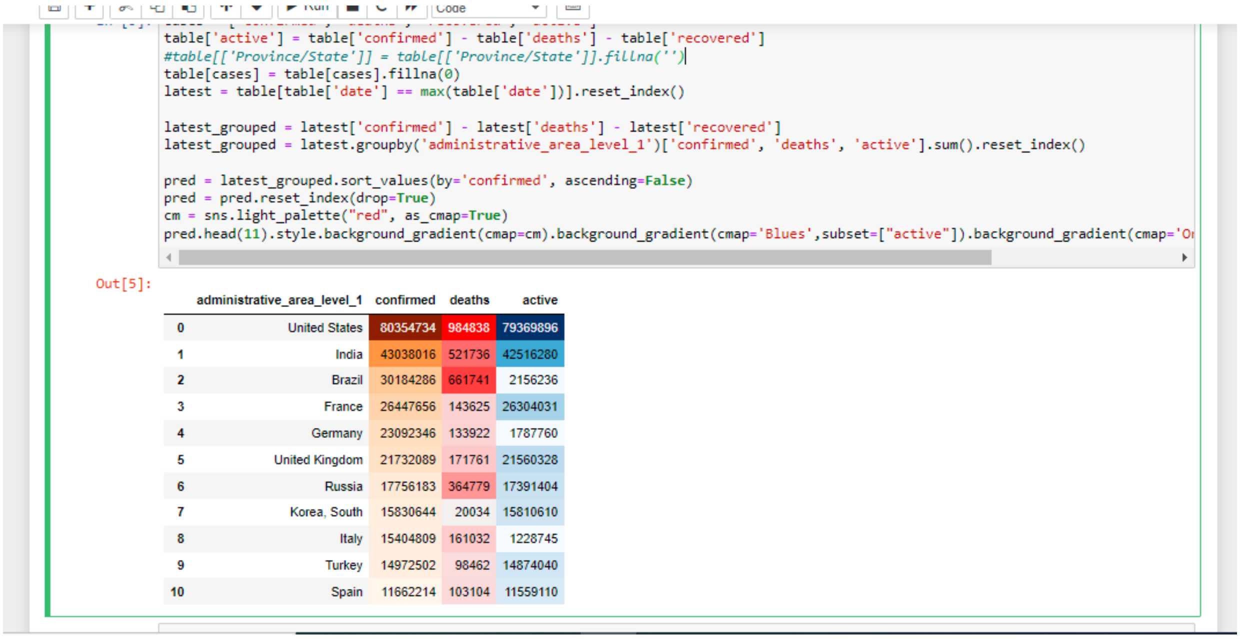


```
2020-01-06    0
2020-01-07    0
```

```
In [39]: df.plot(color='#5D3FD3', figsize=(15,6));
```

```
In [ ]:
```

Columns grouped:

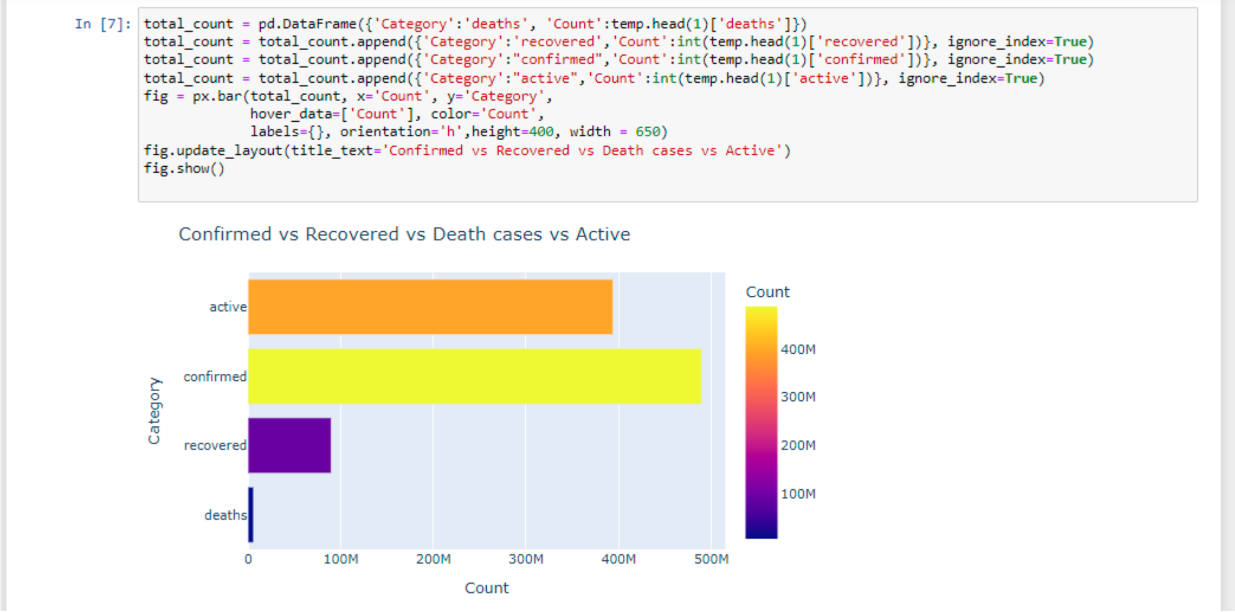


```
table['active'] = table['confirmed'] - table['deaths'] - table['recovered']
#table[['Province/State']] = table[['Province/State']].fillna('')
table[cases] = table[cases].fillna(0)
latest = table[table['date'] == max(table['date'])].reset_index()

latest_grouped = latest['confirmed'] - latest['deaths'] - latest['recovered']
latest_grouped = latest.groupby('administrative_area_level_1')['confirmed', 'deaths', 'active'].sum().reset_index()

pred = latest_grouped.sort_values(by='confirmed', ascending=False)
pred = pred.reset_index(drop=True)
cm = sns.light_palette("red", as_cmap=True)
pred.head(11).style.background_gradient(cmap=cm).background_gradient(cmap='Blues',subset=["active"]).background_gradient(cmap='Or
```

Out[5]:

	administrative_area_level_1	confirmed	deaths	active
0	United States	80354734	984838	79369896
1	India	43038016	521736	42516280
2	Brazil	30184286	661741	2156236
3	France	26447656	143625	26304031
4	Germany	23092346	133922	1787760
5	United Kingdom	21732089	171761	21560328
6	Russia	17756183	364779	17391404
7	Korea, South	15830644	20034	15810610
8	Italy	15404809	161032	1228745
9	Turkey	14972502	98462	14874040
10	Spain	11662214	103104	11559110

Bar chart:



In [7]:

```
total_count = pd.DataFrame({'Category':'deaths', 'Count':temp.head(1)['deaths']})
total_count = total_count.append({'Category':'recovered','Count':int(temp.head(1)['recovered'])}, ignore_index=True)
total_count = total_count.append({'Category':"confirmed",'Count':int(temp.head(1)['confirmed'])}, ignore_index=True)
total_count = total_count.append({'Category':"active",'Count':int(temp.head(1)['active'])}, ignore_index=True)
fig = px.bar(total_count, x='Count', y='Category',
             hover_data=['Count'], color='Count',
             labels={}, orientation='h',height=400, width = 650)
fig.update_layout(title_text='Confirmed vs Recovered vs Death cases vs Active')
fig.show()
```

REFERENCES:

1. Wang J., Liu Y., Wei Y., Xia J., Yu T., Zhang X., Zhang L. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. Lancet. 2020;395(10223):507–513.

2. JA Backer, D Klinkenberg and J. Wallinga, Incubation period of 2019 novel coronavirus (2019-nCoV) infections among travellers from Wuhan, China, 20–28 January 2020.

3. Tolga Ergen and Suleyman Serdar Kozat, "Efficient online learning algorithms based on LSTM neural networks", IEEE transactions on neural networks and learning systems, vol. 29, no. 8, pp. 3772-3783, 2017.

4. H. Li, S.-M. Liu, X.-H. Yu, S.-L. Tang and C.-K. Tang, "Coronavirus disease 2019 (COVID-19): current status and future perspectives", International Journal of Antimicrobial Agents, pp. 105951, 2020.

5. Ameer Sardar, Kwekha Rashid, Heamn N. Abduljabbar & Bilal Alhayani, Applied Nanoscience (2021) “Coronavirus disease (COVID-19) cases analysis using machine-learning applications”